REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 7/01/02	3. REPORT TYPE AND DATES COVERED Final Technical Report

4. TITLE AND SUBTITLE
Modeling and Solution Procedures for Diversity Maximization

5. FUNDING NUMBERS
N00014-00-1-0598

6. AUTHORS
Gary Kochenberger, Fred Glover, Bahram Alidaee, & Keith Womer

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Mississippi, University, MS

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ONR
Dr. Tanja F. Blackstone

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Unlimited

13. ABSTRACT *(Maximum 200 words)*

Many important problems involve selecting a subset from a larger population such that the aggregate diversity of the subset selected is a large as possible. This problem, known as the diversity maximization problem, is known to be NP-hard. As such, it is very challenging from a computational point of view and only very small (toy) problems can be solved to optimality. Accordingly, most research into the important area has focused on various heuristic approaches.

In this research, we report on a new tabu search-based approach to the diversity maximization (MD) problem. We also indicate how additional considerations (constraints) can be folded into the MD model and readily accommodated. Our results are very attractive in terms of both effectiveness and efficiency as we are now quickly solving problems many times larger than previously reported in the literature. Moreover, the solution method we developed for MD has application to a wide variety of other problem areas for which a common thread with MD had not previously been noted in the literature. Thus we are able to greatly leverage the work pioneered here to provide solution procedures for other problems.

14. SUBJECT TERMS
Optimization, Integer Programming, Diversity

15. NUMBER OF PAGES
Nine

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Final Technical Report

GRANT #: N00014-01-1-0598

PRINCIPAL INVESTIGATORS:

Dr. Kieth Womer
Director, Hearin Center for Enterprise Science
School of Business
University of Mississippi

Dr. Gary Kochenberger
School of Business
University of Colorado at Denver &
Hearin Center for Enterprise Science

Dr. Fred Glover
School of Business
University of Colorado at Boulder &
Hearin Center for Enterprise Science

Dr. Bahram Alidaee
School of Businees &
Hearin Center for Enterprise Science
University of Mississippi

INSTITUTIONS:

(see above)

GRANT TITLE:

Modeling and Solutions Procedures for Diversity Maximization

AWARD PERIOD:

01-Apr- 2000 through 31- Mar-2002

OBJECTIVE:

Our main objective in this research was to develop and test a new method for solving diversity maximization problems based on the modern Metaheuristic called Tabu Search. Secondarily, our intent was to investigate the transferability of the procedure for solving MD into other classes of problems. Our results indicate that we were very successful on both accounts.

INTRODUCTION

The diversity maximization problem can be stated as follows. Consider a set of elements $S = \{s_i : i \in N\}$, defined over the index set $\mathrm{N} = \{1,2,...,n\}$, where each element, s_i, has r attributes denoted by s_{ik}, $k \in R = \{1,2,...,r\}$. The objective is to select a subset of size m, where m is strictly less than n, to maximize the *diversity* of the elements chosen.

To express the objective function formally, we associate a measure of diversity d_{ij} with each pair of elements s_i and s_j; that is, d_{ij} is some function of the elements s_{ik} and s_{jk}, $k \in R$, which is selected by the decision maker according to the context. Then the problem can be represented:

$$MaxD : \max\ x_0 = \sum_{i \in N} \sum_{j \in N} d_{ij} x_i x_j = xDx$$

subject to

$$\sum_{j \in N} x_j = m$$

where $d_{jj} = 0$ and x_j is a binary variable denoting whether or not element j is chosen to be a member of the selected subset.

The *MaxD* model is quite general and is capable of representing problems from a wide variety of areas. In spite of the apparent simplicity of its formulation, the model contains an unsuspected capacity for handling multiple considerations of considerable complexity. Included among these are situations where the goal is not merely to select elements that are diverse, but that also satisfy required levels of quality along multiple dimensions.

We begin by sketching some of the general areas of diversity optimization where the maximum diversity concept is critically important.

1. *Environmental Balance*: Ecological systems depend on diversity for survivability. Considerations of diversity maximization are crucial for establishing systems that are viable, robust and balanced.
2. *Medical Treatment*: Combating diseases, both by preventive planning and after the onset of illness, is enhanced by programs that offer more diverse lines of defense in order to combat the broadest spectrum of potential disease causing agents.
3. *Genetic Engineering*: Recombinant DNA and RNA applications yield a richer field of outcomes by designs that generate greater number of alternatives, and where those alternatives, in turn, embody greater diversity in their underlying structure.
4. *Molecular Structure Design*: The quest for improved molecular structures, which affect fields ranging all the way from medicine to metallurgy, depends on finding stable ways to fit molecular shells, and to appropriately position component molecules in available candidate locations. Processes to achieve this have so far been limited by the range of diversity in the elements that are generated and interrelated by standard approaches.

5 *Agricultural Breeding Stocks*: In both animal and plant genetics, the goal of obtaining new varieties by controlled breeding strategies is aided by drawing on breeding stocks with desirable qualities of diversity. Better ways of characterizing and generating subsets of stocks with maximum diversity directly contribute to this goal.

6 *Right Sizing the Firm*: Organizations that need to engage in "downsizing" are at risk of creating critical skill and knowledge gaps when employees with very similar profiles are eliminated. The adverse loss of institutional knowledge can be mitigated by developing a downsizing plan designed to maximize the diversity of those who are retained with the firm.

7 *Composing Jury Panels*: The pursuit of a fair and complete hearing of the evidence brought against a defendant is best served when the case presented is viewed and analyzed from diverse points of view. This ideal is approached by selecting jurors from a pool of qualified citizens with the goal of maximizing the diversity of those chosen.

The applications mentioned above have a common theme of harvesting information from a data base to assist in selecting elements with the greatest variety of characteristics. In such applications, the *MaxD* model serves as an interface between the data base housing the raw data and the decision maker. Implementing the model constitutes an advanced form of data mining by revealing information in the form of optimal solutions – solutions not observable directly from the data in the absence of the model.

MODEL EXTENSIONS

The basic diversity model, *MaxD*, is robust enough to represent a wide range of problems. Nonetheless, further considerations can arise in certain applications that require additional constraints to be added to the model. If these new constraints are linear, they can be accommodated via quadratic penalties within the basic *MaxD* framework; i.e., the further constrained model can be re-cast into the form of *MaxD* and solved by the method illustrated in this paper. The paper by Kochenberger, Alidaee and Amini (1998) discusses such re-formulations in general. We illustrate some useful possibilities of this type by the examples below.

Reformulation Example 1.

Suppose that two elements, which may be quite different from each other on most attributes, are unacceptably close on some critical attribute – so much so that we want to require that not both elements be chosen. Denoting the elements by i and j, we can preclude both from being chosen by imposing the constraint

$$x_i + x_j \leq 1.$$

Such a constraint is not explicitly accommodated by the *MaxD* model. However, we can readily handle the constraint by introducing the penalty term

$$Px_i x_j$$

and subtracting it from the objective function, where P is a suitably chosen positive constant. Since we are maximizing, not both x_i and x_j will receive a value of 1 in an optimal solution.

Denoting by M the set of element pairs that require such mutually exclusive conditions, our modified problem can be written

$$\max x_0 = xDx - P\sum_{(i,j)\in M} x_i x_j = xQx$$

subject to

$$\sum_{j=1}^{n} x_j = m$$

By absorbing the quadratic penalty terms into the matrix D (to produce the matrix Q), we retain the form of the original model, *MaxD*, which enables our tabu search method to be applied without modification. The parameter P must be large enough to force the desired result. Any value greater than an upper bound on the original objective function will clearly work. However, much smaller values have proven successful in practice.

Reformulation Example 2.

The mutually exclusive relationships considered in the preceding example are a special case of a more general type of relationship encountered in a variety of application settings. For example, costs may be attached to the elements to be selected, and a budget limit may be imposed by means of a general linear inequality. Similarly, measures of quality may be associated with the elements, and a linear inequality can be introduced to assure the elements chosen will satisfy an overall quality level. (Multiple measures and inequalities can be introduced to handle definitions of quality of different types.)

A variety of other considerations may likewise lead to further constraints in the form of linear inequalities or equations. In general, linear inequalities over zero-one variables with integer coefficients can be transformed into equations by identifying appropriate bounds on associated slack variables, and then replacing these slack variables by equivalent expansions of zero-on variables.

Whenever the constraining relationships can thus be represented by a system of linear equations in the binary variables, a quadratic penalty (of slightly different construction than the one considered in the previous example) can be employed to incorporate the relationships into the form of the basic *MaxD* model.

To illustrate the approach, consider the further constrained model of the form

$$\max x_0 = xDx$$

subject to

$$\sum_{j=1}^{n} x_j = m$$

$$Ax = b$$

where the equality system *Ax* = *b* represents the additional relationships that need to be taken into account. Taking P, as before, to be a suitably chosen positive penalty, we can re-write the foregoing model as

$$\begin{aligned}\max x_0 &= xDx - P*\left(Ax-b\right)'\left(Ax-b\right)\right)\\ &= xDx + xZx + c\\ &= xQx + c\end{aligned}$$

subject to

$$\sum_{j=1}^{n} x_j = m$$

where the matrix Z and the additive constant c result directly from the matrix multiplication indicated. Thus we are back once more to our basic *MaxD* model, disclosing its broad applicability to DDM problems. This reformulation again affords the opportunity to exploit these problems with our tabu search approach.

APPROACH: (SOLUTION METHODOLOGY)

The *MaxD* model belongs to a class of NP-hard problems, and thus no method is known to exist that is guaranteed to be able to find an optimal solution in "better than exponential" time. In fact, methods that can be proved to converge are unable to find and verify optimal solutions for many problems of realistic sizes within reasonable time limits. Consequently, except for small instances, such problems are preferably approached by heuristic methods rather than exact (theoretically finite) methods.

The literature on *MaxD* contains few papers of a computational nature and the methods reported have been tested on rather small problem instances only. In the original exposition of the model, Kuo, Glover and Dhir (1993) present an equivalent linear mixed integer zero-one formulation of *MaxD* and demonstrate its use on a small example problem. This approach has the advantage of lending itself to readily available, optimal seeking branch and bound algorithms. However, this approach is not viable for problems large enough to be of significant practical interest.

More recently, Ghosh (1996) presents a randomized greedy heuristic for the problem, and Glover, Kuo and Dhir (1998) give several constructive and destructive heuristics. All these methods have been shown to produce high quality solutions on small test problems (30 to 40 variables) where optimal solutions are known. The virtue of these previously reported methods lies in their simplicity. The downside is that for a problem instance of greater dimension, they lack the "intelligence" to navigate a complicated solution space characterized by strong local optima.

The approach we take in this work employs a basic version of tabu search (TS) to guide our search of the solution space. The added search sophistication of TS notably enhances our ability solve problems of much greater dimension and difficulty than is possible by lower level heuristics alone. Our tabu search implementation to solve the *MaxD* model is a variation of the method we have developed and extensively tested for solving the unconstrained binary quadratic program. Detailed descriptions of the method can be found in Glover, Kochenberger and Alidaee (1998) and Glover, Kochenberger, Alidaee, and Amini (1999). Below we give a brief overview of the method.

Tabu Search Overview

Our method is centered around the strategic oscillation approach, which constitutes one of the primary strategies of tabu search. The variant of strategic oscillation we employ alternates between constructive phases that progressively set variables to 1 (whose steps we call "add moves") and destructive phases that progressively set variables to 0 (whose steps we call "drops moves"). To control the underlying search process, we use a memory structure that is updated at *critical events,*
which are identified by conditions that generate a subclass of locally optimal solutions. Solutions corresponding

to critical events are called *critical solutions.* For the maximum diversity problem, we modify the definition of a critical event to stipulate that such an event occurs when an add move during a constructive phase or a drop move during a destructive phase yields a trial solution with exactly m variables equal to 1.

A parameter *span* is used to indicate the amplitude of oscillation about a critical event. We begin with *span* equal to 1 and gradually increase it to some limiting value. For each value of *span*, a series of alternating constructive and destructive phases is executed before progressing to the next value. At the limiting point, *span* is gradually decreased, allowing again for a series of alternating constructive and destructive phases. When *span* reaches a value of 1, a *complete span cycle* has been completed and the next cycle is launched.

Information stored at critical events is used to influence the search process by penalizing potentially attractive add moves (during a constructive phase) and inducing drop moves (during a destructive phase) associated with assignments of values to variables in recent critical solutions. Cumulative critical event information is used to introduce a subtle long term bias into the search process by means of additional penalties and inducements similar to those discussed above. A complete description of the framework for the method is given in Glover, Kochenberger, Alidaee and Amini (1999).

ACCOMPLISHMENTS: (Computational Expereince)

We have successfully applied our method to many problem instances. Here we summarize our results from randomly generated problems of size 100, 300, 500, 1000 and 2000 variables. These problems are substantially larger than those reported earlier in the literature. Prior to this study, the largest problem tested had just 100 variables.

For each problem size, five instances of the problem were considered, each with a different value of m. Our test problems were 100 % dense (off diagonal) and the d_{ij} values were randomly generated between 10 and 50. (Our method is not restricted to require the d_{ij} coefficients to satisfy sign conditions such as nonnegativity, however.) Results from our runs are shown in Table 1.

TABLE 1. Test Problem Results

# vars (n)	M	# TS cycles	Soln Time	Solns Opt?
100	10, 15, 20, 25, 30	20	2 sec (each)	*
300	30, 45, 60, 75, 90	50	15 sec (each)	*
500	50, 75, 100, 125, 150	100	58 sec (each)	*
1000	100, 150, 200, 250, 300	100	194 sec (each)	*
2000	200, 300, 400, 500, 600	200	16 min (each)	*

* Optimal solutions not known

As indicated above, we solved a total of 25 random instances of sizes ranging from 100 to 2000 variables. For each size, five different values of "m" were considered as shown in the table above. For all problems, the q_{ij} values were chosen from U(-50,50). Transformation #1 was used with P = 2n. Our tabu search heuristic was run for a fixed number of cycles, terminating in each case with a feasible solution. Optimal solutions for these problems are not known. However, we have also solved much smaller problems for which optimal solutions are known and in each such case our approach was successful in finding the optimal solution. All runs were made on a Pentium 200 PC.

Prior to this paper, the largest instances reported in the literature were of size n = 100. Our results greatly expand the state of the art and illustrate a solution capability much in excess of that reported by others.

PORTABILITY OF SOLUTION METHOD

The solution method developed and tested here for the maximum diversity problem can, via simple reformulation methods, be applied to a wide variety of other important problem classes. We have, for instance, successfully applied it to graph coloring problems, clique partitioning problems, and a variety of assignment problems. This common solution approach suggest a connectedness among these disparate problems not previously noted (or exploited) in the literature. Further exploration of this common framework is a major part of our on-going research.

CONCLUSIONS:

The diversity maximization model, which can be thought of as a data mining tool (Diversity Data Mining, DDM), makes it possible to identify subsets of populations that maximizes measures of diversity, and has important applications in a wide variety of areas. The *MaxD* diversity model we have identified as a foundation for DDM applications gives a means to extract information in a form and quality not otherwise available.

To give these applications practical significance, we have identified a special solution method based on tabu search. Our computational testing shows that this method quickly obtains high quality solutions to problems of dramatically greater size than previously tested in the literature. Such an ability is crucial for handling problems that arise in real world settings.

In addition, we have shown how extensions of the basic model can readily be reformulated to permit relationships of diverse structure and complexity to be captured within the same model framework. Our introduction of a model for diversity data mining, and the demonstration that it can be solved effectively to yield a useful practical tool, opens the door to studies for assessing the potential of DDM in a wide variety of applications.

Finally, we have indicated that the basic solution procedure designed for *MaxD* can be used to solve other important combinatorial problems. Thus, our work here can be leveraged to yield efficient solution procedures in other areas.

SIGNIFICANCE:

The significance of our work is two-fold. First we have developed and tested a new solution procedure for the diversity maximization problem that dramatically increases the size of problem that can be solved in practice. Secondly, we have shown how the solution methods developed here are portable into other areas given simple re-formulation procedures. This second development is a least as important as important as the